

METHOD AND SYSTEM TO ANALYZE DATA

DESCRIPTION

1. Field of the Invention

The present invention relates to a technology for extracting unique concepts from a large amount of data, particularly to a method and a system for acquiring noteworthy and effective knowledge from such data by comparing concepts to which the same category is assigned.

2. Related Art

Information such as complaints, problems and opinions brought from customers to a manufacturer regarding a certain product has been conventionally stored as document data in some form or another. As this document data is sent from different customers, its contents have great variety. When this data was limited, it was easy to analyze it manually. As comments from many customers such as those via telephone support for products are easily electronified and stored today, collected document data becomes so massive that it is far beyond the range of manual analysis. Accordingly, there have been gradually increasing attempts to visualize the contents of a large amount of documents in various forms to facilitate analysis. By conventional methods, however, it has been processed only to the extent of extracting key words centering on noun phrases and displaying the

09612135-070700

While there have been attempts to analyze text by data mining techniques (such as clustering or analysis of correlation rules) since data mining started to receive attention, conventional methods have most often ended up without acquiring any effective results because the unit of analysis extracted from text was merely a key word represented as a character string.

Thus, an object to be attained by the present invention is to provide a method and a system for acquiring unique concepts from a large amount of data.

2

To attain the above objectives, the present invention provides a method and a system for acquiring effective knowledge by extracting concepts of unique characteristics from a large amount of data containing a textual field.

The present invention comprises a concept extractor and a unique concept extractor. The concept extractor extracts categorized concepts from the data containing text data. The unique concept extractor is a device for extracting unique concepts from the extracted concepts, and extracts in the categorized concepts, of the concepts belonging to the same category, a concept whose statistical characteristic is distinguished beyond a threshold with respect to the set which it belongs.

The concept extractor extracts categorized concepts from nonstylized text by using a morphological analysis making use of a vocabulary dictionary or grammatical knowledge, or ambiguity resolution techniques utilizing a category dictionary. The unique concept extractor finds, of the concepts belonging to the same category, a concept whose statistical characteristic is distinguished beyond a threshold with respect to the set which it belongs (threshold) in respective combinations of categories. In

Further in detail, said concept extractor is comprised of the means for morphologically analyzing said textual part of the data, and based on the results of said morphological analysis, generating clauses of said document data, extracting any key word in said clauses as concepts, applying a category dictionary to said clauses to assign concepts (a replacement expression having a representative meaning of the key word) and a category to a key word therein, analyzing the syntax of a sentence comprising said clauses according to the syntactic tree generation rules, and regarding the key words in said clauses to which concepts and a category were assigned, extracting mutually dependent relationships of the key words in the same sentence and extracting said categorized concepts, namely based on said mutually dependent relationships among the key words, extracting combinations of the categories of the concepts in mutually dependent relationships.

4

BRIEF DESCRIPTION OF THE DRAWINGS

Fig. 2 is a flowchart of the data conversion department.

Fig. 4 is a flowchart of the concept extraction department.

Fig. 6 is an example of display area (1) in the display department.

Fig. 7 is an example of display area (2) in the display department.

Fig. 8 is a flowchart of extraction and display of concepts in display area (1).

Fig. 9 is a flowchart of extraction and display of concepts

in display area (2).

Fig. 10 is a typical example of display in display area (1).

Fig. 11 is a typical example of display in display area (2).

Fig. 12 is a diagram explaining the process of the concept extraction department by using a concrete sentence.

Fig. 13 is an example of a GUI screen containing display area (1) and display area (2).

Fig. 14 is an example of implementation of the hardware used in the present invention.

DESCRIPTION OF THE PREFERRED EMBODIMENTS

Fig. 1 is a block diagram showing the outline of the data analyzing system of the present invention. Moreover, as an embodiment, a data analyzing system that analyzes and extracts unique concepts from the data of inquiries by telephone about a computer product, by way of example, is explained. This data analyzing system converts in advance a large amount of inquiry data 150 sent from customers into an analyzable condition at data conversion department 110 so that it can be mechanically analyzed. From this converted data, data with concepts is extracted by using the category dictionary 170, and a means for searching/detecting unique concepts from the extracted data with concepts is provided.

Further in detail, block 110 in Fig. 1 is a data conversion department that receives input of inquiry data 150 and outputs labeled data 160. It creates data (labeled data) in the form of identically keeping nonstylized data contained in inquiry data 150 and stylized data also contained therein. Here, the stylized data is an easily analyzable data format that primarily comprises two or more pieces of item information, etc., each of said pieces of item information having its start position and allowable number of characters predetermined. On the contrary, nonstylized data is a data format that is difficult to analyze since it is mainly information of variable length is diverse such as natural language text. In addition, block 120 is a concept extraction department that assigns a category to a key word of an input labeled data by using the category dictionary 170, and extracts, among the key words to which a category is assigned, those in mutually dependent relationships in the same sentence as the concepts representing more concrete meanings (labeled data with concepts 180).

Here, concepts means a "key word" with a "category" assigned, and a combination due to mutually dependent relationship of said concepts is further extracted as combined concepts (higher order "concept"). Also, a "label" includes a "category" and a data attribute.

Block 130 is search/characteristic detection department 130 that receives labeled data input with concepts 180 and

The above block 110, 120 and 130 are explained in detail below.

First, examples of inquiry data input in block 110 are as follows.

Call classification: Guidance

-----Example of inquiry data--end-----

Block 110 receives input of the inquiry data 150 as above and converts it into labeled data as follows.

CTQ: Japanese cannot be used on my notebook PC, so I reinstalled the OS. Since then, the modem and the Ethernet

cards cannot be used.

-----Labeled data--end---

Thus, inquiry data including nonstylized data is converted into labeled data as above so as to convert various types of data into the same format. In the above example, TI indicates the title, CT the original inquiry contents, and KW+2byte the type of item. This 2byte indicates the category, Q3 the problem classification, Q4 the call classification, Q2 the response/reaction classification, P3 the solution period, and P4 the call time. Here, a label is information indicating the category of item contents, namely fixed-length information including a category such as "KWM1MT:". Fig. 2 shows a flowchart of the data conversion department. Inquiry data 150 is read in step 210, it is determined whether the data has ended in step 220, and if not, the format is converted in step 230. If the data has ended, conversion is terminated in step 240. For instance, if "Title:" of the inquiry data is found, then it is converted into "TI" and the item contents "Japanese cannot be used on a notebook PC" is recorded beside it. Thus, in the data conversion department, inquiry data is labeled and stylized in a preprocessing stage in order to facilitate analysis of data for concept extraction. It may be easily presumed that such conversion for any data is possible for this trader by changing the conversion rules.

Concept Extraction Department 120

```
-----Example of a category dictionary-----
```

-----Example of a category dictionary-end---

11

merely a character string (for instance, if a character string "Washington" is handled merely as a key word, it cannot be effectively analyzed since it is not clear whether it is a person's name or a place name, while it will have a meaning when a category such as [Person's name] or [Place name] is assigned). As for substantives (words equivalent to nouns), a category is assigned referring to the above category dictionary. As for predicates, the category dictionary is used as with substantives and also categories such as [Problem], [Request] and [Question] are assigned from information of attached words. For instance, the verb "kowareru" is extracted as a concept of "go out of order" belonging to the [Problem] category if the data of (kowareru [verb] go out of order [Problem]) is in the category dictionary, whereas expressions of "cannot..." and "want to..." can be interpreted, without referring to the category dictionary, as [Problem] and [Request] respectively, since it is self-evident, considering that it is data of inquiries by telephone, that they are a problem and a request respectively.

Block 350 is a syntactic tree analysis device for generating, according to the simple syntactic tree generation rules, a syntactic tree of a sentence comprising clauses whose key words name had a category assigned in block 330.

[Block 360 is a mutually dependent relationships extraction device for extracting, among the key words in a clause to

which categories are assigned, those in mutually dependent relationships in the same sentence as the concepts representing more concrete meanings. This block 360 extracts as concepts (labeled data with concepts 370), based on mutually dependent relationships among the key words acquired as a result of a syntactic analysis by syntactic tree analysis device 350, combinations of the categories of the key words in mutually dependent relationships. An example of labeled data with concepts 370 is as follows.

-----Example of labeled data with concepts-----

ID199901010000001

TI Japanese cannot be used on notebook PC

KWM1MT: Product A

KWQ3TC: General guidance

KWQ4TD: Guidance

KWQ2PT: Window service

KWP3SD: 1 day

KWP4CM: 21 minutes

CTQ: Japanese cannot be used on my notebook PC, so I reinstalled the OS. Since then, the modem and the Ethernet cards cannot be used.

KWN1 notebook PC

KWN0 Japanese

KWV2 cannot be used

KWW6 notebook PC... cannot be used

KWN2 OS

KWV6 reinstall

KWWD OS... reinstall

KWN1 MODEM

KWN1 Ethernet card

KWV2 cannot be used

KWW6 modem... cannot be used

KWW6 Ethernet card... cannot be used

-----Example of labeled data with concepts--end---

As mentioned above, labeled data with concepts 370 takes the form of labeled data 160 acquired at the data conversion department with the data extracted at the concept extraction department 120 added so as to be data in the same format as labeled data 160.

Fig. 12 explains the flow of the concept extraction department of the present invention based on an actual sentence. First, if the input sentence "MODEM to Ethernet ga tsuka e nai. (the modem and the Ethernet cards cannot be used)" is entered in step 1210, the sentence is separated into words and a part-of-speech number is assigned to each word by morphological analysis device 310 in step 1220. Thus, the sentence "MODEM to Ethernet ga tsuka e nai. (the modem and the Ethernet cards cannot be used)" is converted as follows. [MODEM, 104][to, 81][Ethernet, 104] [card, 104] [ga, 75][tsuka, 10][e, 44][nai, 51][., 100]

In the above example, they represent as follows. 104...proper noun, 81...case particle "to", 75...case particle "ga", 10...verb stem, 44...adjective subjunctive inflection, 51...negative auxiliary verb "nai",

100...punctuation mark.

Next, clauses are generated in steps 1230 and 1240. In clause generation 1 in step 1230, strings of words of the sentence morphologically analyzed are put together in a clause. A rule of "separating clauses with {81, 75, 100, ...}" is predetermined, and the rule is applied from the beginning of the sentence to separate it by each clause from the beginning. In the case of the above input sentence, there are three clauses from the beginning, and the first words of them are a noun, a noun and verb respectively, which are thus determined to be a substantive phrase, a substantive phrase and an actional phrase respectively. Consequently, the input sentence is converted as follows.

```
{[MODEM, 104][to, 81]}
{[Ethernet, 104] [card, 104] [ga, 75]}
{[tsuka, 10][e, 44][nai, 51][., 100]}
```

Next, clause generation 2 in step 1240 makes every clause of the clauses separated in clause generation 1 a pair of an independent word and an attached word. As for a substantive phrase, if it includes two or more nouns, they are linked in order from the beginning, such as {[Ethernet, 104] [card, 104] → [Ethernet card, 104]}. After that, the part-of-speech code of the independent word is rewritten as N1 representing a general noun phrase. As for an actional phrase, the string of attached words ([e, 44][nai, 51][., 100]) is analyzed, [nai, 51] indicating negative information

```
{[MODEM, N1][to, 81]}
{[Ethernet card, N1] [ga, 75]}
{[tsukaeru, -V1][., 100]}
```

```
(MODEM N1 modem NA)
(Ethernet card N1 Ethernet card NA)
(tsukaeru -V1 tsukaenai VC)
```

```
{[MODEM, NA][to, 81]}
{[Ethernet card, NA] [ga, 75]}
{[tsukaenai, VC][., 100]}
```

16

a sentence comprising clauses to which a category is assigned. The form of the rule of a mutually dependent relationship at this time is (an independent word of the source clause of the mutually dependent relationship, an attached word of the source clause of the mutually dependent relationship, an independent word of the target clause of the mutually dependent relationship, an attached word of the target clause of the mutually dependent relationship). This rule is applied from clause 1{[MODEM, NA][to, 81]} at the beginning of the sentence onward. In general, to the n-th clause, the rule of a mutually dependent relationship is applied for the clauses from n+1-th to the last N-th (n = 1 to N-1). Since there is a rule of (NA, 81, VC, *) in the rule of a mutually dependent relationship, it is determined that there is a mutually dependent relationship between {[MODEM, NA][to, 81]} and {[tsukaenai, VC][., 100]}. * in the rule means that it matches any part of speech or category. This is performed with (n = 1 to N-1) and the clause including information of a mutually dependent relationship is represented as a digraph to convert it into the form of (a clause number of the source clause of the mutually dependent relationship, a clause number of the target clause of the mutually dependent relationship, an independent word, a category, a part-of-speech number of an attached word). Consequently, the input sentence is converted as follows.

(1, 3, "MODEM", NA, 81)
 (2, 3, "Ethernet card", NA, 75)

Moreover, NULL indicates that there is no target mutually dependent relationship.

Eventually, the following concept information was extracted from the sentence of the original document, "MODEM to Ethernet card ga tsuka e nai.".

"Ethernet card...tsuka e nai" "Hardware...problem"

Fig. 4 shows a flowchart of the process of the concept extraction department 120.

In step 420, sentence T in labeled data 160 is divided into morphemes W_0 to W_m . Here, morpheme W is represented by character string w and part of speech p, namely $W = \{w, p\}$. (The above is the process of morphological analysis device.)

Next, in step 430, it is determined whether all the words were converted into clauses, and if so, the process moves on to step 440, and if not, it is determined whether word W_n is an attached word or a punctuation mark in step 432, and if the result is No, word W_n is added to clause P_i in step 434. Here, clause P is a set of one or more consecutive words W, and $P = \{ W^* \} = \{ \{w, p\}^* \}$. And the process returns to step 430. If the result of the determination in step 432 is Yes, word W_n is attached to clause P_i in step 436, and then the next clause is prepared. The process thereafter moves on to step 440. In step 440, it is determined whether all the clauses were processed, and if the process is complete, it moves on to step 450, and if not, in step 442 the clause is converted from the form of $P = \{ \{w, p\}^* \}$ to $P' = \{ \{w_1, p_1\} \{w_2, p_2\} \}$ (here, w_1 is an independent word and w_2 is an attached word). For instance, if $P = \{ [\text{kokusai (international), noun}] [\text{jousei (situation), noun}] [\text{ha, particle}] \}$, the noun phrases are put together in one and it becomes $P' = \{ [\text{kokusai jousei (international situation), noun}] [\text{ha, particle}] \}$. And the process returns to step 440. (The above is the process of clause generation device 320.)

In step 450, it is determined whether dictionary consulting

In step 460, it is determined whether a syntactic tree is completed. If the syntactic tree is completed, the process moves on to step 470. If the syntactic tree is not completed, the process performs a generally implemented

syntactic analysis in step 462, and consequently Pn and Pk are linked. (The above is the process of syntactic analysis device 350.)

In step 470, it is determined whether the extraction of mutually dependent relationships is complete. If not complete, the process moves on to step 472, and based on the rule, extracts any binary relation linked to Pn and registers it in the labeled database with concepts 180. At this time, it refers to mutually dependent relationship rule 474. The rule of mutually dependent relationship rule 474 comprises a set of entries [px py] representing [source mutually dependent relationship category, target mutually dependent relationship category] respectively. For instance, in the case of Pn = {[kikai, hardware][ga, particle]}, Pk = {[kowareru, problem][NULL, NULL]} (n and k are in a mutually dependent relationship), the above rule is used and [hardware, problem] → [kikai, kowareru] is extracted and is registered in labeled database with concepts 180. If the determination in step 470 is Yes, the process is terminated in step 480. (The above is the process of mutually dependent relationship extraction carried out by device 360.)

Search/Characteristic Detection Department 130

Fig. 5 shows a block diagram of search/characteristic extraction department 130. Search/characteristic detection department 130 is comprised of the blocks of input

Instruction analysis device 520 analyzes an instruction received from input (instruction) device 570 to send concepts as a parameter to each device. Input (instruction) device 570 is equivalent to keyboard 6, mouse 7, etc. in Fig. 14, and is used, following a user's instructions to have any desired search and display performed for the data analyzing system. Relative frequency calculation device 530 is a device for calculating relative frequency for the whole or a subset of a document. Relative frequency is calculated here by comparing each concept contained in the whole or arbitrary set X with a set of concepts contained in arbitrary set Y.

22

concepts.

Categorized concept frequency calculation device 550 is a device for acquiring the number of concepts contained in the whole or a subset of a document in each category and in order of frequency according to output of instruction analysis device 520. Examples of output of said device are as follows (in the following examples, INPUT specifies the category, N1 is a category representing [hardware], and OUTPUT is [key word occurrence frequency]).

----Examples of output of categorized key word frequency calculation device----

[INPUT] CATEGORY	N1
[OUTPUT] hard disk	2033
[OUTPUT] monitor	1432
[OUTPUT] printer	1001
[OUTPUT] modem	420
[OUTPUT] scanner	212
[OUTPUT] Ethernet card	143
[OUTPUT] mouse	3

----Examples of output of categorized key word frequency calculation device--end-----

Display department 510 is comprised of a GUI screen including display area (1) shown in Fig. 6 and display area (2) shown in Fig. 7. A user selects various items displayed in display department 510 or enters a parameter for a search or the like by input (instruction) device 570 so that it displays various results (frequency display, search results display, etc.) in display department 510. For instance,

Fig. 8 shows a flowchart of extraction and display of characteristic concepts in display area (1). In step 820, category A and category B are selected. Here, category A and B are axes x and y in display area (1) respectively. Since what is characteristic about element Ax of category A is displayed in a later calculation, the category to be compared is set as B. In step 840, category A is input to device 540, and the concepts contained in category A are

Fig. 9 shows a flowchart of extraction and display of other

frequency is noted and the point where "Product A" intersects with "Slow" is selected (clicked). As shown in Fig. 11, category [Hardware] is checked in display area (2) in a narrowed down state. Then, "hard disk" is in the second highest frequency position and its relative frequency is also high (7.18 times), thus it can be presumed that this product A has a unique problem on its "hard disk". Moreover, when the highest position falls under a product number of this product or the like, so it can easily be ignored.

Fig. 13 shows an example of the most characteristic GUI in the present invention that performs the above steps of operation within one screen. In Fig. 13, category [Month] is selected on the left of display area (2) (not illustrated). Then, the concepts contained in category [Month] are extracted by device 550 in order of frequency. And the relative frequency of the concepts acquired is calculated by device 530, and the concepts contained in category [Month] are displayed on the right of display area (2). Next, "November" is selected on the right of display area (2) (not illustrated). Thus, the data sets are narrowed down by device 540 to "November" of category [Month].

Next, category [Problem] is set as axis X and category [Model name] as axis Y in display area (1). Then, the concepts contained in category [Problem] are extracted by device 550 in order of frequency. Also, the concepts

A floppy disk is inserted into floppy disk drive 20, and this floppy disk, hard disk drive 13 (or a storage media such as CD-ROM 26 or DVD 32) and ROM 14 can store codes or data of a computer program and an operating system for issuing an instruction to CPU and so on in combination with an operating system and implementing the present invention, which is executed by being loaded on memory 4. These codes of a computer program can also be compressed or split for storing on more than one media.

29

Speaker 23 receives via amplifier 22 sound or an audio signal D/A (digital/analog conversion) converted by audio controller 21 so as to output it as sound or voice. Audio controller 21 also makes it possible to A/D (analog/digital) convert audio data received from microphone 24 and incorporate into the system audio data from outside. It is also possible to substitute a voice command for operations of the GUI command department of the present invention by using an application such as ViaVoice (a trademark of IBM Corp.). Furthermore, it is also possible to read aloud displayed search results containing characteristic concepts by using an application such as Home page Reader (a trademark of IBM Corp.).

30

computer, various household electrical appliances such as a TV set with a built-in computer, a game machine with a communication facility, a telephone, a fax, a portable telephone, a PHS and an electronic organizer or a combination of those. However, these components are quoted by way of illustration, and all of them are not essential components of the present invention.

Advantages of the Invention

The present invention provides a method and a system for acquiring effective knowledge from a large amount of data. It also provides an analyzing method and a system by means of GUI of superior usability for acquiring effective knowledge.

00000000-00000000